Minireview

Recent developments and future directions in computational genomics

Sophia Tsoka, Christos A. Ouzounis*

Computational Genomics Group, Research Programme, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK

Received 19 May 2000

Edited by Gianni Cesareni

Abstract Computational genomics is a subfield of computational biology that deals with the analysis of entire genome sequences. Transcending the boundaries of classical sequence analysis, computational genomics exploits the inherent properties of entire genomes by modelling them as systems. We review recent developments in the field, discuss in some detail a number of novel approaches that take into account the genomic context and argue that progress will be made by novel knowledge representation and simulation technologies.

Key words: Genomics; Bioinformatics; Algorithm; Molecular biology database; Genomic context; Systems modelling; Knowledge representation

1. Introduction

Computational biology and bioinformatics have grown from a peripheral discipline to a central field in modern biological science. In the general computational biology arena, genes are identified, translated and compared against the databases, protein functions are tentatively obtained, sequences are clustered into families of related proteins and associated with functional roles within the cell. A number of reviews have appeared over the last years, summarising some of the above activities [1–4].

Computational genomics can be defined as a discipline of computational biology which deals with the analysis of entire genome sequences. But today, computational genomics is much more than mere sequence analysis. Although its roots lie in more traditional bioinformatics methods, there have been significant steps towards a more integral analysis of genome information, including metabolic pathways [5], signalling networks [6], functional classes [7], phylogenetic patterns [8], protein fold types [9] and genome organisation [10]. This increasingly intensified computational approach to genome analysis has generated not only tools for experimental biologists but also interesting scientific results [11].

*Corresponding author. Fax: (44)-1223-494471.
E-mail: ouzounis@ebi.ac.uk

Abbreviations: LIMS, laboratory information management system; DNA, deoxyribonucleic acid; HMM, hidden Markov model; EC, enzyme commission; Database acronyms not included

In this review, we discuss developments in computational genomics and show how this field is becoming a mature discipline, generating both exciting new science and supporting technologies for experimental biology. This review is divided into three parts. First, a brief description of the general methodological approaches relevant to genome sequence analysis is given. Second, we concentrate on some recent developments in the field of computational genomics, which attempt to detect and describe molecular function by exploiting genome structure. Third, we discuss possible future directions in the field of computational genomics with a view to address bottlenecks in molecular function analysis through novel approaches for data representation and integration.

2. The past: from sequences to genomes

One common and useful classification of computational genomics approaches follows the pipeline of large-scale genome analysis employed in various settings, both academic and commercial. These techniques arise from the 'mainstream' bioinformatics activities and mainly focus on genome sequence analysis, for example gene finding, sequence diagnostics, database searching, sequence clustering and functional annotation.

We outline each of the above approaches in turn, refraining from a detailed overview of the vast range of associated problems and we provide reference to recent literature for further reading. We do not address issues of sequencing, quality control, laboratory information management system (LIMS) integration, sequence assembly, or any other aspect of experimentally obtaining genome data. Our virtual pipeline assumes finished deoxyribonucleic acid (DNA) or protein sequences as input data. The engineering aspects of database production, maintenance and updating are only touched upon here, when relevant to the broader discussion. For molecular biology database pointers, the reader is referred elsewhere [12].

2.1. Gene finding

The problem of gene finding in the past few years has mostly focused on gene detection in prokaryotic genomes (bacteria and archaea) [13]. Since the publication of the baker's yeast, worm and fruit fly genomes and in anticipation of the first draft of the human genome, emphasis on successful methodologies for eukaryotic genomes is required. The problem is becoming increasingly thorny [14] due to the vast gaps in terms of phylogenetically related species. The genomes of the worm, fly and now human are isolated in the compendium of species elucidated by genome sequencing projects.

That implies that so-called extrinsic approaches [15] (searching protein databases with the query DNA sequence

0014-5793/00/$20.00
PII: S0014-5793(00)01776-2

for the identification of protein-coding genes) are not as effective as in the case of prokaryotes. Even for prokaryotic genomes, inconsistencies of open reading frame calling abound [16]. The intrinsic approaches [15] of gene detection (predicting genes from first principles such as exon/intron boundary detection) lack the appropriate amount of learning sets for the training of the algorithms [17]. Progress in this area includes the development of hidden Markov model (HMM)-based methods for gene structure that detect more accurately exon/intron boundaries, such as GeneMark [18] and Glimmer [19]. For an elegant, recent review, see also [20].

2.2. Sequence diagnostics

With this general term, we define all approaches that detect sequence features on protein sequences without resorting to searching the database. Evidently, some of the predictive elements may have been derived from databases (such as the propensity of forming trans-membrane segments), but these are not necessarily used directly during the analysis. This step of sequence analysis is fundamental for the characterisation of the query sequences, especially when no similarity to other sequences in the database is readily identifiable. In this category, we cite the detection of coiled-coil [21], trans-membrane [22,23], cellular localisation signal [24,25] and compositionally biased [26–28] regions.

2.3. Database searching

This is probably the most familiar stage of sequence analysis to most biologists, a common activity for many scientists who have sequenced and analysed a gene of interest. The database search stage provides indications not only of family membership (when a set of sequences is identified as being homologous to the query sequence) but also of the possible function of the query sequence, when the homologues have been experimentally characterised and appropriately annotated [29].

However, the noise levels of the database annotations have been increasing since the deposition of genome sequences which are in turn annotated by similarity [30]. Extreme caution is necessary to detect experimentally characterised homologues, usually from a vast number of entries with varying annotation quality. The source of the annotation for a whole family may come from a single sequence and this is not readily obvious. Filters such as low-complexity masking aid in eliminating spurious hits arising due to compositionally biased regions [31].

The area of database searching has seen much growth recently, mostly focusing on providing more sensitive searches. Some of these methods include profile vs. profile methods such as LAMA [32] and HMM-based methods such as Hmmer [33], MAST [34] and SAM-T98 [35]. Interesting work includes the benchmarking of algorithms for their ability to detect weak sequence similarities [36,37].

2.4. Sequence clustering

Further processing of genome information and a stage where quality control of the annotations can take place is the clustering of genes and proteins into families. Clustering of homologous genes has other various applications, but in the context of genome analysis the cross-checking of annotations may be its single most important role [38]. This step allows associated annotations to be cross-checked within families of proteins so that under-predictions, over-predictions and false positives are corrected. Resources in this field include Pfam [39], COGS [40], WIT [41], Protomap [42] and Emotif [43]. One exciting development is the detection of multi-domain proteins which may also provide clues to the function of their single-domain counterparts in complete genomes [44,45] (see below).

Protein fold recognition can also be considered a form of clustering, where target sequences are associated with known fold types. From the genomics perspective, recent developments in this area include the use of sequence threading for the detection of structural fold types within entire genomes [46,47] and the use of correlated mutations for the detection of fold recognition [48].

2.5. Functional annotations

The annotation of sequences is a central aspect of genome analysis and entails the association of sequences with a number of qualitative traits such as enzyme commission (EC) numbers for enzymes, domain structure for multi-domain proteins, molecular roles for proteins of known function, literature pointers and relevant database records such as accession numbers and dates of last update. With such a panoply of analysis methods, our ability to detect function from sequence on the basis of homology to experimentally characterised proteins is continually improving.

However, there are still a number of problems which hamper accurate and, more importantly, consistent functional annotations for genome sequences. First, the transfer of function via homology is a subject of current research [49–52] and no clear-cut rules may immediately apply. Second, the transfer of this information, even when all other criteria are satisfied, crucially depends on the quality of transient database annotations [53] which may be far from satisfactory (no published material on the quality control of curated database annotations is available). Third, the reproducibility of sequence annotations is poor [54–58] and the result is a conundrum of descriptions for genome sequences without a clear consensus.

The best annotations currently available take the form of community-curated databases centred around model organisms, for example EcoCyc [59] for *Escherichia coli*, SGD [60] for *Saccharomyces cerevisiae* and FlyBase [61] for *Drosophila melanogaster*.

2.6. Association with functional roles

This final step in genome annotation is the most important and technically the most challenging of all. This particular aspect of genome analysis is where the whole activity transcends the boundaries of 'classical' sequence analysis and necessitates technology that has yet to be developed. The idea is that the appropriately structured (and potentially formal) function descriptions of gene products can be integrated into systems that represent a general network of cellular processes, including metabolic pathways, transcription activation mechanisms and intracellular control cascades [62].

This area has experienced an unprecedented growth in the last few years. In some ways, this 'role association' can be regarded as a meta-annotation of functional roles of individual genes and an integration step for a comprehensive thesaurus of gene function for a particular species. Unfortunately, standard procedures are not yet in place, and there is great variability both in terms of quantity and quality of the various

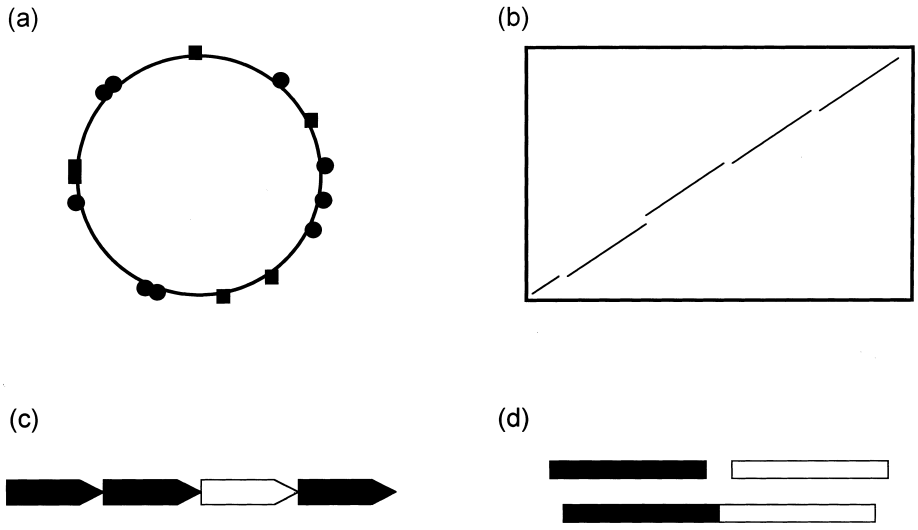

Fig. 1. Pictorial representation of four computational genomics methods. (a) Genome subtraction aims to define species-specific genes. This is achieved by subtracting genes homologous to various elements such as genes orthologous to the species under consideration or phages which are likely to be inserted by horizontal transfer. The method thus detects species-specific genes that can be linked to phenotypic features (represented by squares or circles). (b) Whole-genome alignment for two hypothetical species. Axes indicate genome positions and each point indicates a match between genome sequences. Such genome alignment plots reveal organisational features such as homologous regions or duplications. (c) Functional coupling of gene clusters detects orthologous genes between species which are then used to predict functional networks. The detection of a conserved battery of genes of known function (black arrows) implies that a gene of unknown function (white arrow) may have a related role, on the basis of its presence in the same 'operon'. (d) Schematic representation of fusion analysis. The approach resembles an in silico two-hybrid system and is based on the detection of groups of non-homologous genes in one organism found fused in the corresponding gene in another organism. In the case of genes of unknown function being involved, such associations may be used to infer functional associations.

approaches taken. Metabolic databases [63] have formed a basis upon which other complex categorisation schemes have been developed.

Some of the most successful attempts here include various approaches to metabolic reconstruction, defined as the prediction of the metabolic complement for a species based on the analysis of genome sequence [64–66]. Such systems include EcoCyc [59], KEGG [67] and WIT [41]. Other integrated approaches include GAIA [68] and GO [69].

2.7. Towards computational genomics

A conclusion from the above outline of methodological approaches, and especially the last two ones, is that genome analysis is much more than sequence analysis and includes elements of reaction detection and metabolic reconstruction, which contribute towards a more consistent and reliable integration of functional information [70]. It must be said, however, that biological databases will have to reflect biological reality more closely (for example genome structure, functional networks etc.) before we can ever hope that all this information can be put into genuine use by computational methods.

3. The present: from genomes to systems

In the past few years, advances in the area of computational genomics produced substantial scientific results and established the field as an independent discipline. This challenges the utilitarian, 'toolkit' attitude towards bioinformatics [71,72], and should result in a wider appreciation of this field in the experimental biology community. The wealth of approaches for the analysis of genomes is very extensive and we decided to briefly summarise only some recent developments.

This section concentrates on results that underline the impact of computation on large-scale biology. Most comparative analyses of biological systems today rely on extensive computational approaches such as the ones outlined below. We describe genome subtraction, whole-genome alignment, functional coupling and fusion analysis for the detection of protein interactions. All these approaches have the potential to increase our capacity to propose models of functional networks or associations beyond the threshold achieved by homology-based methods.

These methods depart from traditional computational methods described in the first part, because they primarily depend on the availability of entire genome sequences for the detection of patterns that rely on the organisation and finiteness of genomes. Genome subtraction can only pick out unique genes if the genome sequence is completely known. Whole-genome alignment requires entire genomes, almost by definition. Finally, the precision performance of the last two methods crucially relies on completeness [45]. Thus, all the approaches described next can be defined as computational genomics methods according to our original definition.

3.1. Genome subtraction

Entire genomes allow the detection of 'unique' sequences, genes that are not present anywhere in the database or in the close relatives of the species under investigation (Fig. 1a). These unique sequences may be the key determinants for species-specific phenotypic properties, such as pathogenicity, secondary metabolism properties and the like [73]. These elements are sometimes components of cellular pathways that remain to be discovered and can be interesting drug targets in pathogenic organisms. To identify unique sequences, however, one has to detect equivalent (or 'orthologous') genes,

which are not always easy to define [74]. Despite this shortcoming, this method will be most valuable for the comparison of bacterial strains or other, closely related species. Two interesting studies using this virtual subtraction method have appeared for *Haemophilus influenzae* [73] and *Helicobacter pylori* [75].

3.2. Whole-genome alignment

Another area where technical advances resulted in some deeper understanding of the genome structure and thus function of certain species is whole-genome alignment [10]. Previous systems could not cope with hundreds of kilobases of raw DNA sequence. This advance will facilitate detailed comparisons of genome organisation (revealing single nucleotide polymorphisms, translocations or inserts, repeats and syntenic regions in chromosomes) (Fig. 1b). Another application is strain comparison, reminiscent of genome subtraction. This method has been applied to the comparison of two *Mycobacterium tuberculosis* strains, *Mycoplasma genitalium* and *Mycoplasma pneumoniae* and regions from mouse chromosome 6 against human chromosome 12 [10].

3.3. Functional coupling of gene clusters

Another method that exploits genome structure and organisation is the prediction of functional association of neighbouring genes. It has been observed that certain conserved gene clusters (which may be operons) contain functionally related genes [76–78]. Thus, even for genes of unknown function, there is a possibility to predict their cellular roles [76] or a more specific functional property [78] on the basis of their neighbouring genes (Fig. 1c). Applications include the comparative analysis of two bacterial genomes [76], the comparison of nine bacterial and archaeal genomes to propose physical interactions of gene products [77] and the use of gene clusters from 31 complete genomes to infer functional coupling and reconstruction of metabolic networks [78].

3.4. Fusion analysis

Finally, based on the observation that the homologues of certain genes appear to fuse during the course of biological evolution, this approach attempts to predict functional association and protein interactions on the basis of gene fusion. The methods rely on the assumption that individual 'component' proteins whose homologues are involved in a fused, multi-domain protein must be involved with each other in a protein complex, biochemical pathway or another cellular process [44,45] (Fig. 1d). Detection of false positive predictions by this approach is difficult, mainly due to the lack of extensive experimental information about protein interactions.

3.5. Towards a scientific discipline

The explosion in computational analysis methods for complete genomes brought out not only technologies but also some key scientific results. We are listing some interesting developments that have appeared in the recent literature in the areas of metabolic reconstruction and comparative genomics, using computation alone.

For metabolic reconstruction, examples include the reconstruction of the metabolic networks of *Methanococcus jannaschii* [79], the analysis of the tricarboxylic citric acid cycle across a number of species [80], the characterisation of the known metabolic complement of *E. coli* [81], the distribution of functional classes across the domains of life [82] and the prediction of functional networks in yeast [83].

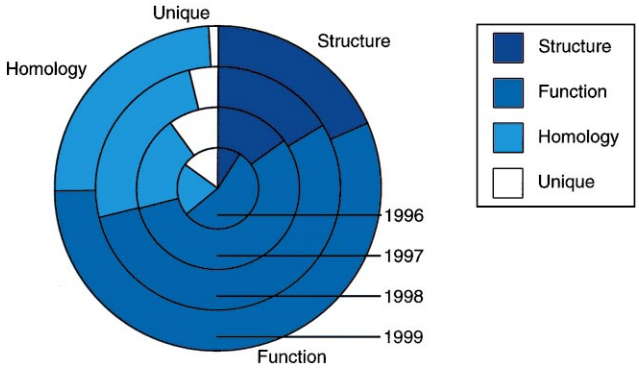

Fig. 2. The 'function bottleneck'. Information clock illustrating the improvement of annotations identified for a given genome over the years 1996–1999. The four levels of annotation range from homologues of known structure (blue) and homologues of known function (marine) to homologues of unknown function (cyan) and unique sequences (white). Note that although structure and homology increased over the years, the function prediction level stalled. Data from the GeneQuiz system, still available at: http://www.ebi.ac.uk/research/cgg/services/.

For comparative genomics, examples include the detection of an archaeal genomic signature [84], the compilation of universal protein families [85], the comparison of three entire eukaryotic genomes [86], the detection of eukaryotic signalling domains in archaea and bacteria [87], the distribution of individual protein families across species [88], the patterns of protein fold usage in microbial genomes [89], the derivation of the universal tree based on enzyme families [90] and the derivation of species relationships based on gene content [91].

Taking a closer look at the properties of entire genome sequences, two things become apparent: first, comparative analysis greatly enhances our abilities to 'predict' and detect molecular function using sequence information and second, the current bottleneck in genomics appears to be the turnover of experimentally obtained novel properties for molecular families of unknown function (Fig. 2). Once the 'function universe' is covered, it may be that computation will acquire a truly central role in biological science.

4. The future: from systems to function

What is to be expected from computational genomics in the near future? As illustrated in the previous sections, our battery of tools is becoming increasingly sophisticated and our ability to detect protein function using computation is generally improving.

However, to resolve the issue of function description and detection, we need to progress from methods mostly derived from traditional sequence analysis that examine genome sequences individually to algorithms and databases that exploit the inherent properties of entire genomes. We are in the process of discovering the constraints that apply to entire genomes so that genomic context can be reflected in our future methods, enhancing the quality of function descriptions.

We argue here that all our approaches towards the elusive goal of predicting function from sequence have to take into account the genomic context and describe molecular function in terms of actions and interactions within the cell. In other

words, our procedures from sequence to function require the development of models that describe cells as systems, using their genetic blueprint, i.e. genome sequence.

4.1. Querying biological databases

It is indisputable that publicly available databanks play a fundamental role in disseminating sequence data to the biological community. However, one of the most important problems of biological data repositories is their archive-like nature. Public databases are designed to store information in an unstructured way, largely in free-text flat-file format without defined object relations. This may help end-users that occasionally browse to retrieve individual entries, but it is very far from making the database amenable to large-scale computation. In this sense, these repositories are not genuine database systems, designed for flexible querying and large-scale data mining.

The query capability for most public databases is fairly limited and mainly consists of keyword-based information retrieval. Consider the following query: 'what are the known protein kinases in the full yeast genome that are involved in cell division, and how many of these have homologues in plants?' The answer requires detailed descriptions of the entities under consideration (in this example: species, e.g. plants; protein classes, e.g. protein kinases; other properties, e.g. homology; cellular processes, e.g. cell division). This query cannot be issued to public databases, in the absence of a formal and detailed classification of biological function, sometimes referred to as an 'ontology'.

Moreover, curation at its present labor-intensive stage has resulted in archives lagging behind completed genome sequencing projects (Fig. 4). To overcome the curation hurdle, more efficient solutions must be developed. We believe that the only way to transfer all our biological knowledge from the amorphous 'textome' available as free text in journals, books and periodicals is to build robust and customised natural language processing systems that will be in wide use in the very near future (for an overview of this subfield, see also [92]). In addition to browsing and querying, an important application of this technology is its large-scale deployment for the creation and true integration of new-generation molecular biology and genomics databases.

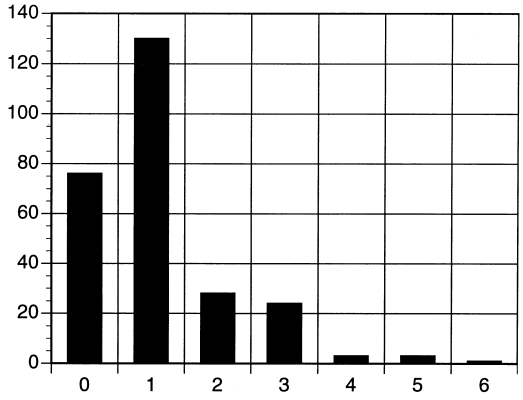

Fig. 3. The distribution of EC numbers (y-axis) versus the number of pathways they have been assigned to (x-axis) for the genome of *M. jannaschii*. Data from GenePOOL (Ouzounis et al., unpublished).

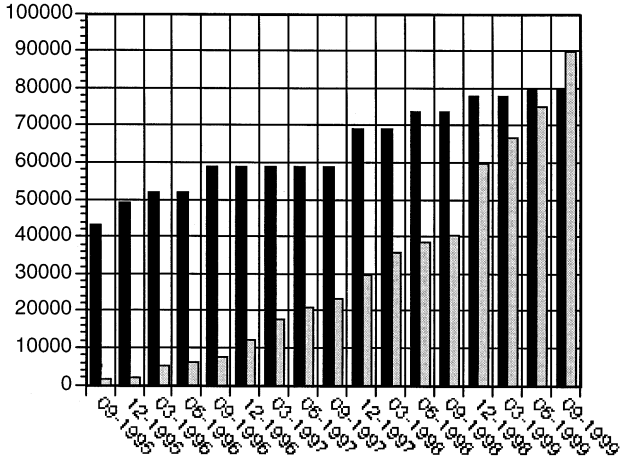

Fig. 4. Comparison of the growth of the Swiss-Prot database entries (black) and the available protein sequences for complete genomes (grey), for corresponding Swiss-Prot release dates.

4.2. Classifications of biological function

To accurately describe function, biologists have to agree on a common vocabulary and classification of molecular roles for all genes and proteins. This is an immense task, made much more difficult by its nature as a community project. The amount of data is significant, but finite. It is the complexity of the information that makes this task daunting.

The issue of data complexity can be tackled by portable 'ontology' designs, exact specifications of various conceptualisations for a given domain. Systems that allow the exchange and development of these schemes have been imported from other areas of science and engineering (for a web-based knowledge server, see also [93]). This strategy is one way of dealing with highly complex, qualitatively rich features of specific domains of discourse. Without elaborating further, we should note that a number of systems that have addressed this issue already exist in genomics. Such systems attempt to classify and further process various aspects of molecular function in terms of general hierarchies for genome sequence and biochemical pathways (e.g. EcoCyc [59]), ribosome structure and function (e.g. RiboWeb [94]), cellular processes and function categories (e.g. GO [69], see also geneontology.org) and generalised functional classes (e.g. GeneQuiz [29]). Although the latter are simple, general and also automatically derived [95], they have yet to be widely accepted. One reason may be the clash of opinions on the definition of functional classes, and the relatively restricted utility of a high-precision but low-coverage classification of protein functions [96].

To showcase the utility of such systems, we briefly describe here some of our own work. We have devised a simple 'ontology' design, called GenePOOL, that stores structured information about all computationally derived annotations for the genome of *M. jannaschii* (Ouzounis et al., unpublished). It is based on a flat-file exchange format called Genome AnnoTatiOn System (GATOS) (developed in collaboration with Peter Karp, AI Center, SRI International). GenePOOL allows queries that not only improve the consistency of the data but also allow the discovery of novel patterns. A query that cannot be issued to public databases for *M. jannaschii* is: 'what is the distribution of EC numbers across metabolic pathways?' (Fig. 3). This type of analysis can be applied to automatic metabolic reconstruction based on the detection of

reaction information (Tsoka and Ouzounis, in preparation). More work along this direction is needed, so that all elements of molecular function become amenable to large-scale computation.

4.3. Structural genomics

This review would be incomplete without a note on the structural genomics initiatives that have been proposed in the recent literature [97,98]. The idea here is to massively solve the structure of all proteins for a given genome. Thus, the function of all proteins will be determined by the sheer knowledge of their structure, a well-known motto in structural biology [99]. Significant progress has already been made in terms of assigning structural homologues to proteins of known function for a number of completely sequenced species [100] (Fig. 2). It should be noted, however, that some recent claims for structural genomics may be slightly overstated [101,102]. The issue of how structure can contribute to the prediction of function is still an open question, despite the invaluable amount of information that can be extracted from structural analysis and comparison.

4.4. Other developments

Without further elaborating, we would like to mention a few other developments which we believe are shaping the future of computational genomics towards an even richer, more complex and stimulating field than it has ever been. These include DNA chip data analysis [103], gene expression analysis [104] (see Brazma and Vilo, in this issue), genetic regulation networks [105], simulation environments for whole-cell modelling [106], detection of regulatory networks [107], modelling gene regulation dynamics [108], complex modelling of metabolic networks [109,110] and tissue-specific data-driven resources, e.g. the human brain [111].

4.5. Towards biological simulation

We have argued here that truly integrated functional annotations for gene products should reflect the corresponding biological properties of the molecules under consideration. Computational genomics faces a formidable task; all biological knowledge has to be properly mapped, assembled, classified, encoded, represented, modelled, updated and maintained with all the components and dimensions of molecular function accessible for computation. Only then true function description will have been achieved. In the future, we should be able not only to obtain all possible molecular functions and relations instantly and reliably, but also to simulate the full network of molecular interactions for a cell, tissue, organ, organism or population, all the way down to the molecular level.

Acknowledgements: We thank Richard Coulson for critical reading of the manuscript, other members of the Computational Genomics Group at the EBI for discussions, and numerous colleagues for comments. Due to space limitations and the vast amount of literature, we apologise to colleagues whose work has not been adequately cited in this review. C.O. acknowledges support from the EMBL, the Medical Research Council (UK), the European Commission (DGXII – Science, Research and Development) and IBM Research.

References

[1] Overbeek, R., Larsen, N., Smith, W., Maltsev, N. and Selkov, E. (1997) Gene 191, GC1–GC9.
[2] Brutlag, D.L. (1998) Curr. Opin. Microbiol. 1, 340–345.
[3] Danchin, A. (1999) Curr. Opin. Struct. Biol. 9, 363–367.
[4] Skolnick, J. and Fetrow, J.S. (2000) Trends Biotechnol. 18, 34–39.
[5] Karp, P.D., Krummenacker, M., Paley, S. and Wagg, J. (1999) Trends Biotechnol. 17, 275–281.
[6] Takai-Igarashi, T., Nadaoka, Y. and Kaminuma, T. (1998) J. Comput. Biol. 5, 747–754.
[7] Riley, M. (1998) Curr. Opin. Struct. Biol. 8, 388–392.
[8] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Proc. Natl. Acad. Sci. USA 96, 4285–4288.
[9] Gerstein, M. (1997) J. Mol. Biol. 274, 562–576.
[10] Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. (1999) Nucleic Acids Res. 27, 2369–2376.
[11] Andrade, M.A. and Sander, C. (1997) Curr. Opin. Biotechnol. 8, 675–683.
[12] Kreil, D.P. and Etzold, T. (1999) Trends Biochem. Sci. 24, 155–157.
[13] McIninch, J.D., Hayes, W.S. and Borodovsky, M. (1996) ISMB 4, 165–175.
[14] Thanaraj, T.A. (2000) Nucleic Acids Res. 28, 744–754.
[15] Borodovsky, M., Rudd, K.E. and Koonin, E.V. (1994) Nucleic Acids Res. 22, 4756–4767.
[16] Raghavan, S. and Ouzounis, C.A. (1999) Nucleic Acids Res. 27, 4405–4408.
[17] Burset, M. and Guigo, R. (1996) Genomics 34, 353–367.
[18] Lukashin, A.V. and Borodovsky, M. (1998) Nucleic Acids Res. 26, 1107–1115.
[19] Salzberg, S.L., Delcher, A.L., Kasif, S. and White, O. (1998) Nucleic Acids Res. 26, 544–548.
[20] Stormo, G.D. (2000) Genome Res. 10, 394–397.
[21] Lupas, A., Van Dyke, M. and Stock, J. (1991) Science 252, 1162–1164.
[22] Kihara, D., Shimizu, T. and Kanehisa, M. (1998) Protein Eng. 11, 961–970.
[23] Pasquier, C., Promponas, V.J., Palaios, G.A., Hamodrakas, J.S. and Hamodrakas, S.J. (1999) Protein Eng. 12, 381–385.
[24] Nakai, K. and Horton, P. (1999) Trends Biochem. Sci. 24, 34–36.
[25] Nielsen, H., Brunak, S. and von Heijne, G. (1999) Protein Eng. 12, 3–9.
[26] Wootton, J.C. (1994) Comput. Chem. 18, 269–285.
[27] Wootton, J.C. and Federhen, S. (1996) Methods Enzymol. 266, 554–571.
[28] Promponas, V.J., Enright, A.J., Tsoka, S., Kreil, D.P., Leroy, C., Hamodrakas, S., Sander, C. and Ouzounis, C.A. (2000) Bioinformatics (in press).
[29] Andrade, M.A. et al. (1999) Bioinformatics 15, 391–412.
[30] Karp, P.D. (1998) Bioinformatics 14, 753–754.
[31] Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Nat. Genet. 6, 119–129.
[32] Henikoff, S., Henikoff, J.G. and Pietrokovski, S. (1999) Bioinformatics 15, 471–479.
[33] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D. and Sonnhammer, E.L. (1999) Nucleic Acids Res. 27, 260–262.
[34] Bailey, T.L. and Gribskov, M. (1998) J. Comput. Biol. 5, 211–221.
[35] Karplus, K., Barrett, C. and Hughey, R. (1998) Bioinformatics 14, 846–856.
[36] Muller, A., MacCallum, R.M. and Sternberg, M.J. (1999) J. Mol. Biol. 293, 1257–1271.
[37] Henikoff, S., Pietrokovski, S. and Henikoff, J.G. (1998) Nucleic Acids Res. 26, 309–312.
[38] Enright, A.J. and Ouzounis, C.A. (2000) Bioinformatics 16 (in press).
[39] Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) Nucleic Acids Res. 28, 263–266.
[40] Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) Science 278, 631–637.
[41] Overbeek, R. et al. (2000) Nucleic Acids Res. 28, 123–125.
[42] Yona, G., Linial, N. and Linial, M. (1999) Proteins 37, 360–378.
[43] Nevill-Manning, C.G., Wu, T.D. and Brutlag, D.L. (1998) Proc. Natl. Acad. Sci. USA 95, 5865–5871.
[44] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Science 285, 751–753.

[45] Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Nature 402, 86–90.
[46] Jones, D.T. (1999) J. Mol. Biol. 287, 797–815.
[47] Fetrow, J.S. and Skolnick, J. (1998) J. Mol. Biol. 281, 949–968.
[48] Olmea, O., Rost, B. and Valencia, A. (1999) J. Mol. Biol. 293, 1221–1239.
[49] Shah, I. and Hunter, L. (1997) ISMB 5, 276–283.
[50] des Jardins, M., Karp, P.D., Krummenacker, M., Lee, T.J. and Ouzounis, C.A. (1997) ISMB 5, 92–99.
[51] Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) J. Mol. Biol. 297, 233–249.
[52] Hegyi, H. and Gerstein, M. (1999) J. Mol. Biol. 288, 147–164.
[53] Wheelan, S.J. and Boguski, M.S. (1998) Genome Res. 8, 168–169.
[54] Brenner, S.E. (1999) Trends Genet. 15, 132–133.
[55] Andrade, M., Casari, G., de Daruvar, A., Sander, C., Schneider, R., Tamames, J., Valencia, A. and Ouzounis, C. (1997) Comput. Appl. Biosci. 13, 481–483.
[56] Tsoka, S., Promponas, V. and Ouzounis, C.A. (1999) FEBS Lett. 451, 354–355.
[57] Pallen, M., Wren, B. and Parkhill, J. (1999) Mol. Microbiol. 34, 195.
[58] Kyrpides, N.C. and Ouzounis, C.A. (1999) Mol. Microbiol. 32, 886–887.
[59] Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M. and Pellegrini-Toole, A. (2000) Nucleic Acids Res. 28, 56–59.
[60] Cherry, J.M. et al. (1998) Nucleic Acids Res. 26, 73–79.
[61] Gelbart, W.M. et al. (1997) Nucleic Acids Res. 25, 63–66.
[62] Karp, P.D. and Paley, S. (1996) J. Comput. Biol. 3, 191–212.
[63] Karp, P.D. (1998) Trends Biochem. Sci. 23, 114–116.
[64] Gaasterland, T. and Selkov, E. (1995) ISMB 3, 127–135.
[65] Bono, H., Ogata, H., Goto, S. and Kanehisa, M. (1998) Genome Res. 8, 203–210.
[66] Karp, P.D., Ouzounis, C. and Paley, S. (1996) ISMB 4, 116–124.
[67] Kanehisa, M. and Goto, S. (2000) Nucleic Acids Res. 28, 27–30.
[68] Bailey Jr., L.C., Fischer, S., Schug, J., Crabtree, J., Gibson, M. and Overton, G.C. (1998) Genome Res. 8, 234–250.
[69] Ashburner, M. et al. (2000) Nat. Genet. 25, 25–29.
[70] Galperin, M.Y. and Brenner, S.E. (1998) Trends Genet. 14, 332–333.
[71] Boguski, M.S. (1994) Curr. Opin. Genet. Dev. 4, 383–388.
[72] Benton, D. (1996) Trends Biotechnol. 14, 261–272.
[73] Huynen, M.A., Diaz-Lazcoz, Y. and Bork, P. (1997) Trends Genet. 13, 389–390.
[74] Ouzounis, C. (1999) Trends Genet. 15, 445.
[75] Huynen, M., Dandekar, T. and Bork, P. (1998) FEBS Lett. 426, 1–5.
[76] Tamames, J., Casari, G., Ouzounis, C. and Valencia, A. (1997) J. Mol. Evol. 44, 66–73.
[77] Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Trends Biochem. Sci. 23, 324–328.
[78] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) Proc. Natl. Acad. Sci. USA 96, 2896–2901.
[79] Selkov, E., Maltsev, N., Olsen, G.J., Overbeek, R. and Whitman, W.B. (1997) Gene 197, GC11–GC26.
[80] Huynen, M.A., Dandekar, T. and Bork, P. (1999) Trends Microbiol. 7, 281–291.
[81] Ouzounis, C.A. and Karp, P.D. (2000) Genome Res. 10, 568–576.
[82] Andrade, M.A., Ouzounis, C., Sander, C., Tamames, J. and Valencia, A. (1999) J. Mol. Evol. 49, 551–557.
[83] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) Nature 402, 83–86.
[84] Graham, D.E., Overbeek, R., Olsen, G.J. and Woese, C.R. (2000) Proc. Natl. Acad. Sci. USA 97, 3304–3308.
[85] Kyrpides, N., Overbeek, R. and Ouzounis, C. (1999) J. Mol. Evol. 49, 413–423.
[86] Rubin, G.M. et al. (2000) Science 287, 2204–2215.
[87] Ponting, C.P., Aravind, L., Schultz, J., Bork, P. and Koonin, E.V. (1999) J. Mol. Biol. 289, 729–745.
[88] Tomii, K. and Kanehisa, M. (1998) Genome Res. 8, 1048–1059.
[89] Gerstein, M. (1998) Proteins 33, 518–534.
[90] Doolittle, W.F. (1999) Science 284, 2124–2129.
[91] Snel, B., Bork, P. and Huynen, M.A. (1999) Nat. Genet. 21, 108–110.
[92] Thomas, J., Milward, D., Ouzounis, C., Pulman, S. and Carroll, M. (2000) PSB 5, 538–549.
[93] Abernethy, N.F., Wu, J.J., Hewett, M. and Altman, R.B. (1999) IEEE Intell. Syst. 14, 79–85.
[94] Chen, R.O., Felciano, R. and Altman, R.B. (1997) ISMB 5, 84–87.
[95] Ouzounis, C., Casari, G., Sander, C., Tamames, J. and Valencia, A. (1996) Trends Biotechnol. 14, 280–285.
[96] Tamames, J., Ouzounis, C., Casari, G., Sander, C. and Valencia, A. (1998) Bioinformatics 14, 542–543.
[97] Burley, S.K. et al. (1999) Nat. Genet. 23, 151–157.
[98] Brenner, S.E. and Levitt, M. (2000) Protein Sci. 9, 197–200.
[99] Orengo, C.A., Todd, A.E. and Thornton, J.M. (1999) Curr. Opin. Struct. Biol. 9, 374–382.
[100] Teichmann, S.A., Chothia, C. and Gerstein, M. (1999) Curr. Opin. Struct. Biol. 9, 390–399.
[101] Eisenstein, E. et al. (2000) Curr. Opin. Biotechnol. 11, 25–30.
[102] Shapiro, L. and Harris, T. (2000) Curr. Opin. Biotechnol. 11, 31–35.
[103] Gerhold, D., Rushmore, T. and Caskey, C.T. (1999) Trends Biochem. Sci. 24, 168–173.
[104] Bassett Jr., D.E., Eisen, M.B. and Boguski, M.S. (1999) Nat. Genet. 21, 51–55.
[105] Wagner, A. (1997) Nucleic Acids Res. 25, 3594–3604.
[106] Tomita, M. et al. (1999) Bioinformatics 15, 72–84.
[107] VanBogelen, R.A., Greis, K.D., Blumenthal, R.M., Tani, T.H. and Matthews, R.G. (1999) Trends Microbiol. 7, 320–328.
[108] McAdams, H.H. and Arkin, A. (2000) Curr. Biol. 10, R318–R320.
[109] Simpson, T.W., Follstad, B.D. and Stephanopoulos, G. (1999) J. Biotechnol. 71, 207–223.
[110] Schuster, S., Fell, D.A. and Dandekar, T. (2000) Nat. Biotechnol. 18, 326–332.
[111] Pietu, G. et al. (1999) Genome Res. 9, 195–209.